Bacterial protein sorting: experimental and computational approaches

Grasso, Stefano

DOI:
10.33612/diss.150510580

CHAPTER 1

GENERAL INTRODUCTION AND SCOPE OF THE THESIS

Introduction

Biotechnological exploitation of living (micro-)organisms has started some millennia ago[1], probably as a cause or as a consequence of the development of agriculture and the domestication of animals. The earliest biotechnological products were prototypes of bread[2], beer[3], and later wine[4,5], based on the exploitation of the fermentation process typical of yeasts. Similarly, also the first dairy products, such as yogurt and cheese, can be considered amongst the oldest biotechnological commodities, since for their production enzymes, mainly proteases, are needed. Rennet, the enzyme mix sourced from mammalian stomachs to produce cheese, was the first food ingredient to be replaced by biotechnologically produced chymosin, which was approved by the U.S. Food and Drug Administration in 1991[6]. Yet the market started a few years earlier with the production of enzymes for industrial processes, in particular in the textile industry[7].

The global market of industrial enzymes has been growing ever since, and it is expected to reach the size of 7 billion $ in 2021[7]. Due to this fast growth of the market, combined with the needs for more diverse and cheaper enzymes, many efforts to improve the production in terms of quality and efficiency were undertaken in the last two decades[8].

Microorganism are the favourite source and production tool for industrial enzymes due to their availability, easiness of manipulation and fast growth rates[8]. Bacteria are widely used, but yeast and filamentous fungi are widely employed as well. Because of their broad usage and the fundamental research questions they raise, bacterial protein production and sorting pathways have received much attention over many years. In fact, in order to make enzymes more profitable, and to expand the current list of industrial enzymes, biochemical and genetic tools have been employed to both understand and hack how proteins are produced and where they are transported, within or outside the cell[7,8].

In the last decade, a plethora of computational tools was developed with the aim of supporting and improving metabolic and protein engineering, involving a multitude of different approaches[9,10]. Computational tools may be based on very different types of algorithms and may have different specificities, but they all share the goal of reducing the amount of experimental work and, thus, both time and expenses, by *in* silico simulating and predicting *in vivo* processes.

In the present dissertation, the main focus will be placed on Gram-positive bacteria and their protein sorting and secretion pathways, using *Bacillus subtilis* as a model organism. In particular the classical secretion pathway will be addressed, which is commonly used to obtain secreted biotechnological products in order to ease downstream recovery and the subsequent purification of proteinaceous products. Additionally, computational tools related to protein sorting and subcellular protein localization (SCL) will be discussed. A specific discussion will be dedicated to the question how bacterial cells select and determine which proteins are to be secreted and how computational tools can help in reproducing and predicting the related bacterial behaviour.

Protein sorting in Gram-positive bacteria

Before discussing how proteins are targeted to the different cellular compartments, it is necessary to determine and define these compartments. Unfortunately, multiple layers of semantic issues have confounded the terminology that is often used[11]. Gram-positive bacteria were traditionally identified based on the retention of a chemical stain, named the Gram stain after the name of its inventor, in the outer part of the cell envelope, which is now known as the cell wall[12]. In contrast, the outer membrane of Gram-negative bacteria does not retain the crystal violet-iodine complex used for the Gram staining. Upon discovery of the cellular structures of different bacteria, the Gram-staining was linked to the presence of a single or double membrane, dividing bacteria into Gram-positive and Gram-negative species, based on differences in their membrane enclosure. Nevertheless, taxonomically speaking, it has become more and more challenging to separately cluster these two groups[13]. In the past, when phylogenetic analyses based on genome sequences were not yet possible, the equivalency of the terms 'Gram-positive = Firmicutes (including Actinobacteria) = monoderm', and 'Gram-negative = Gracilicutes = diderm' was widely accepted[14]. Nowadays, with a better understanding of the ultra-structures of bacteria, and improved phylogenetic analyses, this (subdi)vision has become obsolete[13]. Nonetheless, it is widely retained in scientific publications, which may lead to misunderstandings. In the present thesis, the term Gram-positive will be used to refer to monoderm Firmicutes only, as exemplified by the model organism *B. subtilis*, unless stated otherwise. The Firmicutes have a homogenous and well-conserved cellular structure, allowing a generalization of the protein sorting pathways within them.

As graphically represented in Figure 1 Gram-positive bacteria possess a fairly simple cellular envelope, consisting of a plasma membrane and a thick peptidoglycan layer (i.e. the cell wall). Within the cell, proteins can be targeted to four main compartments: the cytoplasm, the plasma membrane, the cell wall and the extracellular milieu. Given that bacterial proteins are ribosomally synthesized in the cytoplasm, all of them start their journey in this compartment and are subsequently targeted or sorted to their final destination. Additionally, in order to reach the extracellular milieu (i.e. the furthest destination from the place of synthesis), proteins need to transiently cross the plasma membrane and the cell wall[11,15,16]. In some cases, more specificity can be given to the localization of proteins, and this will be discussed separately for each SCL. Moreover, it must be noted that proteins may also have multiple or alternative SCLs, which may depend on internal or external stimuli.

From cytoplasm to plasma membrane

As the journey of all proteins starts in the cytoplasm, the journey of most cytosolic proteins will be quite short, essentially ending with their synthesis or assembly into protein

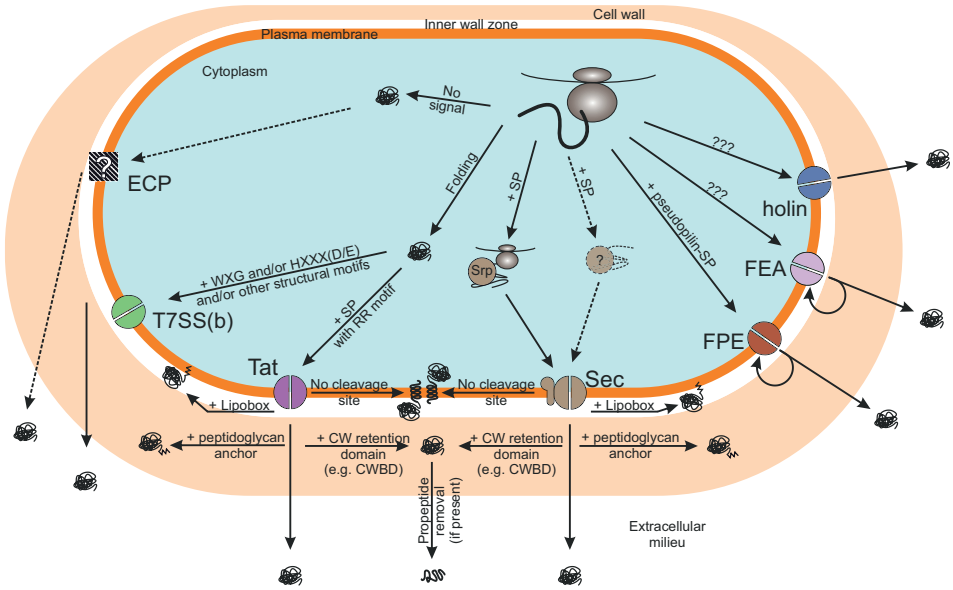

Figure 1. Summary of the known protein secretion pathways in Firmicutes. Cartoon representing a Firmicute bacterium and its different secretion pathways. On the plasma membrane (dark orange) are represented the different translocation apparatuses. In the central position is indicated the general secretory pathway (Sec), which translocates the majority of proteins either co-translationally or post-translationally across the membrane. Proteins are targeted to the Sec translocon via the signal recognition particle (SRP) or other less well-defined chaperones. The sorting signal is the canonical signal peptide (SP). The substrates of Sec are lipoproteins, integral membrane proteins, cell wall proteins and extracellular proteins. To the left of Sec is indicated the twin-arginine translocation (Tat) pathway, which secretes proteins in a folded state, driven by a canonical signal peptide containing a twin-arginine (RR) motif in its N-region. The known substrates of Tat are integral membrane proteins, cell wall proteins and extracellular proteins. Minor pathways for protein secretion, presented clockwise from right to left include: a) holin-mediated secretion for which the sorting signals and mechanisms are not yet properly elucidated; b) the flagellar export apparatus (FEA) that assembles flagella and other motility organelles, and for which the sorting signals are not fully elucidated; c) the fimbrilin-protein exporter (FPE) that is used to secrete and assemble fimbrial, pseudopilus and competence proteins, and for which the sorting signal is the pseudopilin signal peptide. On the left are indicated in clock-wise order: a) the type VII secretion system (T7SS), present mostly in its functional but incomplete T7SSb form, for which the sorting signal seems to be a structural motif that has not been fully elucidated; and b) the non-canonical secretion events that lead to the appearance of extracellular cytoplasmic proteins (ECPs) in the extracellular milieu, and whose secretion rationale, mechanism and signals are presently unknown. The fate of many proteins translocated across the membrane is to become extracellular proteins. However, proteins exported via Sec or Tat may also be: i) retained in the plasma membrane as a lipoprotein through lipidation of the conserved Cys residue in the lipobox; ii) retained in the cell wall either by covalent attachment to the peptidoglycan, or by interaction of specific retention motifs with the cell wall. Importantly, also transmembrane proteins are translocated through the Sec machinery, but instead of being completely translocated, they are laterally sorted from the translocon and in most cases miss a signal peptidase cleavage site. Of note, some proteins possess a Pro-peptide that supports the post-translocational folding process and keeps the enzyme inactive prior and during membrane translocation. The Pros are usually cleaved, and subsequently degraded, once the protein has reached the cell wall or extracellular milieu.

complexes. Instead, all other proteins need to be targeted towards the plasma membrane as a first step. This can generally take place either co-translationally or post-translationally. The vast majority of membrane proteins appears to be co-translationally targeted to the membrane, while for secreted and cell wall-associated (CW) proteins both mechanisms may be employed. Additionally, two fundamentally different secretion pathways co-exist in many bacteria: 1) the general secretion (Sec) pathway, which secretes most proteins co- or post-translationally in an unfolded state; and 2) the twin-arginine translocation (Tat) pathway, which secretes proteins exclusively in a folded state and thus only post-translationally[16–20].

At this point a premise is necessary: most of the targeting and translocation mechanisms have only been studied in *Escherichia coli* and are thus relatively well understood in Gram-negative bacteria. As a consequence, various aspects of protein translocation still remain to be clarified in *B. subtilis*, although they may be very similar to their counterparts in *E. coli*. In particular, in Gram-negative bacteria, proteins can be routed to the Sec pathway post-translationally with the aid of SecB, a chaperone that helps maintaining them in an unfolded state. A homologue of SecB is absent from *B. subtilis*, but a possible analogous protein, the chaperone CsaA, has been identified several years ago. Nevertheless, it is still debated to what extent proteins are co- or post-translationally secreted via the Sec pathway in *B. subtilis*[19,21,22].

Once the nascent protein emerges from the ribosomal exit tunnel, it can be recognized by the signal recognition particle (SRP), a riboprotein complex consisting of an RNA molecule of 271 nucleotides named scRNA, the Ffh protein, and the HBsu protein (also known as HupA or Hbs)[21]. SRP is able to recognize a hydrophobic region in the nascent peptide and weakly bind to it. This hydrophobic region can either be the signal peptide (SP) necessary to translocate proteins across the plasma membrane, i.e. to secrete them, or the first hydrophobic α-helix of a trans-membrane (TM) protein[16,19]. Once the nascent chain is bound by SRP, it is co-translationally targeted with the cognate ribosome, to FtsY. FtsY is a peripheral membrane protein that acts as a receptor of SRP and mediates its docking with the Sec translocase[20]. The translocase complex is composed of a motor, the ATPase SecA, which provides the energy for translocation[22]; the SecYEG channel spanning the plasma membrane and being conserved also in Eukaryotes (Sec61)[20]; and the membrane-integrated chaperone SecDF that employs the proton-motive force to enhance protein translocation[23–25].

At this point the journey of integral membrane proteins and those proteins fully translocated across the membrane diverges. After a continuous channel between the ribosomal exit tunnel and the SecYEG translocon has been formed, the polypeptide chains of integral membrane proteins will be completed and laterally inserted into the plasma membrane[20,26]. During this step, also another protein, YidC, may facilitate the lateral insertion and release of proteins into the membrane. The bacterial YidC insertase was first

characterized in *E. coli*. In *B. subtilis*, two homologues of YidC are present: SpoIIIJ (YidC1) and YqjG (YidC2). The first, SpoIIIJ, seems to be the main YidC, but it is not essential and it is involved in sporulation. Notably, only a depletion of both proteins is lethal[18,27,28]. Importantly, studies in *E. coli* have shown that YidC can also autonomously insert (i.e. without the involvement of SecYEG) small membrane proteins with only one or two TM helices into the membrane, but the mechanism is still not completely understood[17,27,29].

Membrane crossing: canonical Sec-mediated translocation

All non-cytosolic and non-membrane proteins that are targeted to the plasma membrane by SRP or alternative chaperones are doomed to be translocated to the extracytoplasmic side. Once the nascent or already completed polypeptide chain is docked to the SecYEG translocon, SecA dimers bound to the translocon will take over and start providing the necessary energy to push the soon-to-be-secreted protein through the translocation channel. To do so, SecA will go through the following cycle of reactions: 1) binding of ATP to SecA will occur prior to the binding of SecA to SecYEG; 2) SecA will subsequently hydrolyse the bound ATP molecule, giving freedom of movement to the (nascent) polypeptide chain, thereby allowing a portion of the polypeptide to slide trough the SecYEG channel; 3) at this point SecA can either dissociate from the SecYEG complex and bind a new molecule of ATP to start the cycle over, or use the energy generated by the dissociation of ADP to push further the polypeptide chain[15,19,30,31]. It must be underlined that during step 2) multiple conformational changes occur, both in SecA to release the secretory polypeptide and in SecYEG, where the channel slightly opens to allow the polypeptide chain to slide through. Although the overall process has been studied in great detail, the exact mechanism used by SecA to drive protein secretion is still not completely clear. As reviewed in detail by Collinson[32], two main models of the mechanism employed by SecA have been formulated: a power stroke/diffusional hybrid and a diffusional ratchet model. In both models the energy is provided through SecA-mediated ATP hydrolysis, together with the proton-motive-force. Possibly, also an optimal involvement of chaperones to keep the nascent chain unfolded, combined with protein folding of the secreted peptide on the extracytoplasmic side of the membrane, will help in preventing a backward movement of the translocating polypeptide[32].

It must be remarked that during the early steps of translocation, the SP acquires a "reversed" hairpin-like position in the membrane, with its N-terminus pointing toward the cytosol. Subsequently, the remaining portion of the secreted protein will unloop and pass through the SecYEG channel. Upon unlooping of the SP, its C-region emerges from the plane of the plasma membrane and a type I signal peptidase (SPase) will cleave the SP from the mature protein. This cleavage releases the latter into the extracytoplasmic space, and leaves the SP within the membrane. In *B. subtilis* five chromosomally-encoded SPases are known, namely SipS, SipT, SipU, SipV, and SipW; additionally, a plasmid-encoded SPase,

SipP, has been detected[30]. Lastly, in order to avoid accumulation of SPs within the membrane, they are degraded by a signal peptide peptidase (SPPase)[33]. Different membrane-associated proteases have been invoked in the process of signal peptide degradation, in particular SppA and RasP[34]. However, SppA has probably only a minor role in this process, if any[35].

Membrane crossing: Other pathways

The Sec pathway described in the previous section is regarded as the classical or canonical protein secretion route in Gram-positive bacteria. However, variants of this pathway have been identified in *B. subtilis* and other organisms. In addition, dedicated pathways exist for the export of specific groups of proteins from the cytoplasm, e.g. flagellar proteins.

The following pathways play no or only a marginal role in the context of industrial protein production. Yet, they are important when trying to understand bacterial sorting mechanisms in their completeness and, thus, for protein subcellular localization (SCL) prediction, which will be discussed in the second half of this chapter.

SecA2

Some pathogenic Gram-positive bacteria, both Firmicutes and Actinobacteria, possess two SecA proteins, namely SecA1 and SecA2. While SecA1 exerts the "normal" function of SecA and is, thus, involved in the translocation of most proteins, SecA2 is involved in the secretion of only a subset of proteins, often virulence factors[36,37]. Despite SecA2 being a paralogue of SecA, it is often smaller in size and, interestingly, it is not conserved among Gram-positive bacteria. This suggests that it probably evolved independently among the various species[37,38]. While the necessity of a SecA paralogue is not fully understood, nor its precise mechanism, the existence of at least two functional SecA2 pathways has been demonstrated so far. The first one is called the SecA2-only system, and it is most likely associated to the canonical SecYEG translocon. Conversely, the second SecA2 pathway involves an accessory SecY protein, called SecY2, and functions independently from SecYEG[37,38]. Remarkably, it has been challenging to understand what types of proteins are secreted via the SecA2 pathways, how they are recognized, and why they need a dedicated secretion apparatus. Studies on SecA2 from various genera of Gram-positive bacteria showed a lot of peculiarities[39]. However, if SecA2 evolved independently multiple times, it may well be that it has acquired different functions and mechanisms in different genera.

Lipoproteins and SPase type II

Another variation in the canonical Sec pathway is responsible for the export and lipidation of lipoproteins. These proteins are translocated across the plasma membrane and then covalently attached to it, thus localizing on the extracytoplasmic side of the membrane[15,40].

Firstly, lipoproteins are targeted to the secretion machinery, as described above for other proteins, through the Sec pathway. Once being translocated to the extracytoplasmic

side of the membrane, the prolipoprotein diacylglyceryl transferase (Lgt), catalyses the transfer of a diacylglycerol molecule to the conserved Cys in position +1 of the mature protein (i.e. the first amino acid after the cleavage of the SP occurs; see the section on 'Signal Peptide (SP) types'). This anchors the protein to the plasma membrane via the acquired lipidic moiety. Subsequently, a type II SPase, namely LspA in *B. subtilis*, which recognizes a different consensus sequences compared to type I SPases, is responsible for cleaving the SP. In most Firmicutes, the lipoprotein sorting pathway stops here. However, in pathogenic *Staphylococcus* species and Actinobacteria a further step, shared with Gram-negative bacteria, is present. In these species, after the SP cleavage performed by SPase II, a phospholipid/apolipoprotein transacylase, is responsible for also N-acylating the Cys in position +1[41]. In *Staphylococcus aureus*, this N-acylation is catalysed by the LnsAB system and it serves to avoid Tol-like receptor 2-mediated detection of the pathogen. In certain Actinobacteria and Gram-negative bacteria, N-acylation of the +1 Cys residue is a conserved signal for translocating the lipoprotein to the outer membrane, despite the different structures of the respective outer membranes[40,42,43]. It must be remarked here that lipoproteins may also be exported through the Tat pathway (see the next section), both in Gram-negative bacteria and Actinobacteria, suggesting that both the Sec and Tat export pathways can transport lipoprotein precursors that are subsequently lipidated and processed by SPase II. Nonetheless, in Firmicutes no evidence for Tat-secreted lipoproteins has been found so far[42]. Furthermore, many lipoproteins produced by *B. subtilis* and other Firmicutes may end up in the extracellular space due to secondary proteolytic removal of the N-terminally acylated Cys residue[15,40,44].

Twin-arginine translocation (Tat) pathway

The Tat pathway is considered as the other major protein secretion pathway, which is known to secrete many proteins in Actinobacteria[45]. Nevertheless, in Firmicutes only a few cargo proteins have been identified[46,47]. The Tat pathway secretes proteins that are already fully folded and even complexed or associated with their respective co-factors in the cytosol, in contrast to the canonical Sec pathway whose cargo proteins are folded after translocation[46–48]. Similar to the Sec pathway, the Tat machinery and related mechanisms are highly conserved, and present both in Gram-positive and Gram-negative bacteria, as well as in Archaea and in the thylakoid membranes of chloroplasts. Despite its conservation, the Tat pathway was found to be very variable in terms of accessory components and compatibility, even among closely related organisms[46,47,49].

Tat systems are usually composed of a docking complex and a pore complex. The docking complex is needed to recognize and bind the SP of cargo proteins and it is formed by two proteins, a TatA-like protein which has a single membrane-spanning helix, and a TatC protein which has 6 TM helices. In Gram-negative bacteria, the docking complex is formed by TatB, which belongs to the family of TatA-like proteins, and TatC. On the

contrary, *B. subtilis* possesses a minimal set of Tat proteins, encompassing only TatA and TatC, but no TatB. Remarkably though, there are two copies of each component, encoded by the *tatAd tatCd* and *tatAy tatCy* operons. Within this context, both TatAd and TatAy can play the equivalent role of *E. coli* TatB, and form a docking complex with TatCd or TatCy, respectively. Once the docking complex interacted with the SP, the TatC component is responsible for inserting the cargo protein into the plasma membrane. In turn, the just-formed complex, composed of Tat proteins and the cargo, recruits more TatA proteins necessary for the formation of the 'pore complex'. Lastly, the cargo protein is translocated to the extracytoplasmic side of the plasma membrane. To date, the precise details of this step are not yet elucidated, but it either involves the formation of a channel that may vary in size, or local weakening of the membrane[46–48]. It is interesting to note that in *B. subtilis* both the TatAdCd and TatAyCy complexes are sufficiently complete to be fully functional. The main difference found so far between the two complexes lies in their specificity for particular cargo proteins, where the phosphodiesterase PhoD is specifically translocated by TatAdCd, while the Dyp-type peroxidase EfeB (YwbN), the Rieske protein QcrA and the metallophosphoesterase YkuE are specifically translocated by TatAyCy[46–48].

Surprisingly, in *B. subtilis* a third copy of TatA, called TatAc, was detected, but its role is not completely clear as it is not essential for secretion. In fact, despite being able to form complexes with both TatCd and TatCy, it is not able to replace either TatAd or TatAy, but it can still support protein secretion by TatAyCy[50].

Type VII(b) secretion system

Traditionally, secretion systems have been numbered with ordinals only in Gram-negative bacteria, ranging from type I to type VI. Nevertheless, when a novel secretion system was discovered in *Mycobacterium tuberculosis*, which belongs to the Actinobacteria, the assignment of a proper name posed challenging. Initially, based on the name of the main secreted factor, ESAT-6 (early secreted antigen target 6; also called EsxA), the system was named ESX. Eventually, *M. tuberculosis* was shown to possess five of such systems named ESX-1 to ESX-5, which are involved in the secretion of virulence factors. Subsequently, as *M. tuberculosis* is a diderm organism, the novel system was also referred to as the type VII secretion system (T7SS)[51–53].

Although it was initially believed that the T7SS would be restricted to *Mycobacteria*, this system is also active in other Gram-positive bacteria, both Actinobacteria and Firmicutes. Particularly, in Firmicutes the T7SS is present as a simpler variant that does not necessarily seem to be associated with virulence, as exemplified by its presence in non-pathogenic species, such as *B. subtilis*. This minimal T7SS was therefore named T7SSb. An older name for this secretion system in Firmicutes is WXG100 secretion system (Wss), which was derived from the name of a class of cargo proteins[51–56].

The ESX-1, ESX-3 and ESX-5 systems from *M. tuberculosis* are the best

characterized secretion systems of this type. Nevertheless, the roles of many components of this secretion machinery have not been elucidated yet. In Firmicutes, the best studied T7SSb is the one of *S. aureus*, but similar to the situation in *B. subtilis*, some components of this machinery still need to be identified[51,53,56].

Unfortunately, also the cargo proteins of the T7SS are still to be comprehended. It is known that, overall, the systems of this type secrete proteins belonging to many different protein families, including the WXG100, LXG, DUF2563, DUF2580, PE, PPE, Esx and Esp families, all belonging to the EsxAB clan protein superfamily (PFAM: CL0352)[56–58]. These proteins possess some conserved motifs such as the Trp-X-Gly motif, which is present in all T7SS cargo proteins and led to the name WXG100 family, and the Pro-Glu or Pro-Pro-Glu motifs that are specific for the PE and PPE families, respectively. An additional motif, present at the C-terminus of cargo proteins and apparently necessary for their secretion is the H-X-X-X-Asp/Glu-X-X-h-X-X-X-H motif ('H' stands for highly conserved hydrophobic residue, while 'h' for a less conserved one). Of note, in *B. subtilis* only one cargo protein of the T7SS is known, which is secreted as a dimer[54]. This is the YukE protein of unknown function[53].

Non-canonical secretion

Despite the many secretion systems described, there are still secreted proteins that are not translocated across the plasma membrane via one of the aforementioned pathways. In fact, flagellar proteins are exported via a dedicated machinery called the flagellar export apparatus (FEA)[11,15,59], while fimbriae, pseudopili and competence proteins are translocated via a fimbrilin-protein exporter (FPE)[11,30]. An additional secretion system exists for the export of phage-derived proteins, which is named holin, based on its capability to form pores within the bacterial membrane[11,15].

Even taking into account these three additional secretion systems, the membrane translocation of various proteins experimentally detected on the outside of the cell can currently not be attributed to any known secretion system. Most remarkably, many of these proteins have known cytosolic functions. Different hypotheses have been proposed for the mechanisms by which these 'extracellular cytoplasmic proteins' (ECPs) are secreted, and their possible extracellular functions. These range from cell lysis to a still undetected secretion machinery or vesicle-mediated secretion[60–63]. For sure, the most surprising feature of these proteins is the, apparent, lack of a 'secretion' signal that distinguishes them from other cytosolic proteins. In *B. subtilis* it was shown that the amount of non-canonically secreted proteins increases upon the successive deletion of genes for secreted proteases, suggesting that the detectable accumulation of ECPs is directly controlled by proteolysis[64]. In this respect, it is noteworthy that secreted proteases of *B. subtilis* control the activity of autolysins, indicating that the release of ECPs from the cell is to some extent related to lysis[64].

Cell-wall retention signals

Resuming the journey of proteins towards the extracellular milieu, only CW and extracellular proteins are left to be discussed. First of all, it is important to underline that both classes of proteins may be translocated across the plasma membrane in any of the above-described ways. Secondly, all proteins crossing the plasma membrane can eventually become secreted proteins. In essence, some of them are retained within the CW for a certain amount of time and, if no interaction between a particular protein and the CW occurs, it will diffuse into the extracellular milieu becoming an extracellular protein[65,66].

There are two main ways to retain proteins within the CW, namely through covalent attachment to the CW peptidoglycan, or through non-covalent bonds. In the first case, proteins to be attached to the CW, possess both a SP at the N-terminus, necessary to route it out of the cytoplasm, and a conserved motif at the C-terminus. This conserved C-terminal signal is composed of the consensus sequence Leu-Pro-X-Thr-Gly (LPXTG), a hydrophobic domain and a positively charged domain (the most C-terminal part). This signal is recognized by sortases, a family of transpeptidases, whose role is to cleave the sorting signal between the Thr and Gly residues, and subsequently to covalently bind the Thr residue to the peptidoglycan[15,66,67]. With time it has become clear that multiple sortases are present in Gram-positive bacteria. Sortase A (SrtA) is the most common and representative with the highest number of target proteins, i.e. those with an exact LPXTG consensus sequence. Sortase classes B, C and D are also present among both Firmicutes and Actinobacteria, and they recognize slightly different consensus sequences, such as NP[Q/K]TN, NQPTN, LPXTA, or LAXTG[65–67].

While the covalent attachment of proteins to peptidoglycan is mostly understood as it is a fairly homogeneous process, retention of proteins via non-covalent bonds is definitely less clear and it can be achieved via multiple and different domains. Generally, non-covalently CW-bound proteins possess, within the retention domain, specific motifs that, once folded, can bind to the peptidoglycan or somehow interact with it. Examples of these domains are the cell wall binding domains 1 and 2 (CWBD1 and CWDB2), the lysin motif domain (LysM), the GW module (composed of Gly-Trp dipeptides), the S-layer homology domain (SLHD), the peptidoglycan-binding domain 1 (PBD1), the WXL domain (from the Trp-X-Leu motif), and the clostridial hydrophobic domain (ChW). Each of these binding domains has a specific structure and mode of action. For instance, some domains can simply bind peptidoglycan, while others need to be present in tandem or even in multiple repeats[65,66]. Some domains and motifs have been identified due to their presence in many CW proteins. However, it remains to be proven for some of them that they are actually sufficient for peptidoglycan binding, an example being the SH3b (src Homology-3 bacterial) domain[65].

Other domains are often associated with retention in the CW, for instance the

NLPC/P60 domain or the N-acetylmuramoyl-L-alanine amidase domain. However, these are domains with cell wall-modifying enzymatic activities and cannot be considered as proper CW retention signals. In fact, they will bind their CW-derived substrate molecules regardless of their cellular localization or, in case no site to be processed is found, they will not be retained at all[65].

Signal Peptide (SP) types

As already briefly mentioned above, in order to determine the pathway that a protein will follow, a signal must be embedded within its amino acid sequence. Such signals are most often present at the N-terminus of proteins, but a few exceptions exist.

Sec-SPs

Classical Sec-type SPs were the first to be identified, possibly due to their relatively high abundance compared to other sorting signals and their peculiar structure. In fact, they can be virtually divided into three distinct and specific regions: 1) the N-region, approximately 5-6 amino acid residues long, is characterized by the presence of positively charged residues; 2) the H-region, about 15 residues long, is highly hydrophobic and adopts an α-helical structure that will facilitate insertion into the plasma membrane; and 3) the C-region that is usually fairly short (i.e. approximately as long as the N-region or less) and includes a cleavable consensus sequence. Of note, between the end of the H-region and the beginning of the C-region, SPs usually contain a helix-breaking residue. This Pro or Gly residue breaks the α-helix of the H-region, allowing for a less structured C-region that thus becomes accessible to proteases for cleavage[15,30,68].

Despite the common structure of Sec-SPs, they have a little conserved amino acid sequence. This, together with the fact that the secretion efficiency provided by each SP differs, depending on the mature protein it is fused to[69,70], has limited the comprehension of what are the most relevant characteristics of SPs. As extensively reviewed[68], in order to build a comprehensive model of protein secretion, many SP features have been investigated singularly, and even machine learning (ML) models were trained to recognize SPs[71]. Nonetheless, little light has been shed so far on the combined features that determine the SP efficiency. Many characteristics are in fact known to be important, e.g. charge, hydrophobicity, length and the consensus sequence for SPase cleavage, but the respective impact on protein secretion could not be quantified yet. This lack of knowledge and understanding has hampered the *in silico* design of efficient SPs, although a recent advancement was achieved through a ML model able to generate protein-specific SPs, resulting to be functional in 48% of cases[72].

Tat-SPs

Very similar in overall structure to Sec-SPs, Tat-SPs have as their main characteristic the

conserved N-terminal S-R-R-x-F-L-K motif, with x being a polar amino acid. This motif, which encompasses two arginine residues has in fact given the name to the whole pathway. Furthermore, the Tat-SPs tend to be slightly longer compared to the Sec-SPs and, despite presenting slightly different statistics for charge and hydrophobicity, they retain the same tripartite structure and consensus sequence for SPase cleavage[15,30,46,47].

It must be remarked that it is difficult to precisely identify Tat-secreted proteins due to two main factors. Firstly, there is a high degree of similarity between the Sec- and Tat-SPs, which makes prediction prone to high false-positive rates. In fact, this similarity can lead a protein to either of the two secretion machineries. Accordingly, Tat-SPs often contain a so-called Sec-avoidance signal in the form of a positively charged residue at the end of the H-region[73]. Secondly, because Tat-secreted proteins are exported almost exclusively in a folded state, it may even happen that they are secreted after the quaternary structure is already formed. Consequently, there may be 'hitchhiker' proteins that are translocated through the Tat translocon, because they are part of a protein complex and not because of a specific sorting signal[46,47].

Lipoprotein-SPs

Similar to Sec- and Tat-SPs, also lipoprotein- (Lipo-) SPs possess a tripartite structure with comparable characteristics. In addition to being shorter than the other two SP types, Lipo-SPs present their main difference in the consensus sequence of the cleavage site, which is recognized by type II SPases. This consensus sequence, known also as the lipobox, corresponds to L-(A/S)-(A/G)-C, where C is the first residue of the mature protein. This Cys residue is strictly conserved as it is needed for lipidation[15,30].

Pseudopilin-SPs

As exemplified by *B. subtilis* competence proteins that assemble into a so-called pseudopilus, the pseudopilin-SPs have a completely different structure compared to Sec-, Tat- and Lipo-SPs, with an N-region that is not specifically charged, followed by a hydrophobic H-region. Interestingly, the pseudopilin SPase cleavage site, with consensus sequence K-G-F, is positioned between the N- and H-regions. Upon processing, the +1 Phe residue is methylated. In *B. subtilis*, the cleavage and methylation are carried out by ComC and the subsequent pseudopilin translocation and assembly follows a dedicated pathway[15,30].

T7SS-SPs

For a long time, the sorting signals of proteins secreted via the T7SS have been investigated, and it was only recently determined that, rather than a sequence motif, the signal is presented as a tertiary structure. T7SS cargo proteins are often secreted as dimers, where each monomer possesses two α-helices separated by the W-X-G consensus sequence, resulting in a helix-turn-helix structure. Interestingly, larger T7SS cargo proteins may be

secreted as monomers that form a similar four-helix bundle by themselves. Being common to all T7SS substrates, the bundle of four α-helices seems to be recognized by the secretion machinery, thereby serving as the T7SS sorting signal. Additionally, after the helix-turn-helix structure, and located at the C-terminus, there is the aforementioned conserved H-X-X-X-Asp/Glu-X-X-h-X-X-X-H sequence, which also seems to be involved in binding the secretion machinery[53,54,74,75].

Pro-peptides and pro-regions

The term Pro-peptide (Pro) designates a region that is often found between the SP and the mature protein, and that is proteolytically removed after secretion. The main role of Pros, which are found usually in enzymes, such as proteases, is to help and catalyse protein folding. Additionally, upon folding, the Pro may keep the enzyme inactive. Such a mechanism is necessary to avoid cellular damage caused, for instance, by extracellular proteases that could become active in the cytosol. Consistent with this idea, it has been shown that certain Pros can work both in *cis* and in *trans*, so both when being associated with the mature protein and when the Pro and mature protein are present as two separate molecules. No general structure for Pros has been identified so far, which is in agreement with the fact that Pros must adopt different structures and conformations for each different class of cognate proteins[76,77].

It has been proposed that Pros may enhance the secretion levels of heterologously produced proteins[78,79]. However, no specific mechanism has been described so far. Most likely, it is not the Pro itself that has an influence on secretion. Instead, it was shown that the first 5 to 15 (and up to 30) residues immediately after the cleavage site (also referred to as pro-region) can modulate, and specifically increase, protein secretion levels[68,80].

Protein sorting-related prediction

The subcellular localization (SCL) of proteins is usually experimentally determined as part of their functional annotation. Unfortunately, however, this is both time-consuming, expensive, and often impractical due to the high number of variables that should be tested. For such reasons, the past three decades have seen an explosion of bioinformatics tools, many of which are devoted to predicting protein characteristics and properties with the amino acid sequence as the sole input.

SP prediction

SP prediction has a long history (extensively reviewed in [81]), with the first manual prediction procedure developed in 1983 and based "simply" on the length of the uncharged region and the maximal hydrophobicity[82]. With time, more complex and accurate methods were developed, first based on features, then on position-weight matrices, and lastly ML algorithms, such as artificial neural networks, hidden Markov models, and support vector

machines[81]. Eventually, deep learning methods for SP predictions were developed[83,84]. The most famous tool is SignalP, now at the 5th version, but many other effective tools exist as well, such as PrediSi[85], Phobius[86] or SPEPlip[87], and DeepSig[84].

SP prediction consists of two main parts: the detection of the SP itself, and the determination of the SPase cleavage site (i.e. the N-terminus of the mature protein). While combining the detection of both improves the overall prediction of SPs, different tools may vary considerably in one of the two aspects, e.g. a predictor can be accurate in the cleavage site determination, but not in the SP detection. A third aspect is the distinction of different types of SPs, specifically Sec-SPs, Tat-SPs and Lipo-SPs. As discussed above, these three types of SPs are fairly similar, with only a specific twin-arginine motif in the N-region for Tat-SPs and the lipobox at the cleavage site for Lipo-SPs that distinguish them from Sec-SPs. In order to distinguish SP types, different programs were developed. Examples for the prediction of Tat-SPs are TatP[88], TatFind[89,90] and PRED-TAT[91], while examples for the prediction of Lipo-SPs are LipoP[92], SPEPlip[87] and PRED-LIPO[93]. Of note, SPEPlip is able to discriminate between Sec-SPs and Lipo-SPs. More recently, also SignalP introduced this option. With SignalP 5.0[83] it is now possible to simultaneously discriminate all three types of SPs and their cleavage sites. For the sake of completeness, it should also be mentioned that signatures for Sec-SPs, Tat-SPs and Lipo-SPs exist in all major databases, e.g. Interpro[94], Pfam[57], PROSITE[95] and TIGRfam[96].

Notably, all of the mentioned tools based on ML algorithms were trained on relatively small numbers of SPs (in the order of a few hundreds to a couple of thousands) as derived from *E. coli* and *B. subtilis* for Gram-negative and Gram-positive bacteria respectively, and a few related well-characterized organisms. Consequently, predictions will be biased towards the detection of SPs resembling those in the applied training sets, thereby potentially limiting the detection of SPs from more distantly related organisms or SPs with outlying characteristics. Such a bias, which decreases with the availability of more experimental data, must always be taken into account to avoid self-fulfilling prophecies. In particular, the missed detection of novel ("non-standard") SPs may be caused by a missed prediction which, in turn, limits the possibility to improve SP predictions.

Another important consideration is that the mere presence of a SP is sometimes not enough to determine the fate of a protein, set aside for lipoproteins that always localize to the so-called inner wall zone of Gram-positive bacteria[97], unless they are liberated from the membrane by a secondary cleavage step[15,40]. In fact, after translocation initiated by a Sec-SP or Tat-SP, the protein may be retained in the CW or reach the extracellular milieu. Conversely, the non-detection of a SP does not mean that the protein is not translocated across the plasma membrane, as there are many secretion pathways that are poorly understood and whose cognate sorting signals are not yet recognised. This will be discussed in more detail in the following section.

Other sorting and retention signal predictions

In addition to the many SP predictors, a limited number of other sorting and/or retention signal predictors exists. The latter category includes predictors that are dedicated to the detection of TM helices, which immediately identifies integral membrane proteins. Initially, such predictors were based on the hydrophobicity of consecutive residues, but nowadays all tools employ some sort of learning algorithm. Notable examples of the latter category are TMHMM[98], HMMTOP[99], MEMSAT3[100] and OCTOPUS[101]. Interestingly, in order to improve their results, some tools combine both SP and TM helix predictors. For instance, this is the case with SPOCTOPUS[102], or Phobius[86] and PolyPhobius[103]. While the task of predicting TM helices may seem easier compared to the prediction of SPs, the exact determination of where TM helices start and end is still partially inaccurate, often leading to different numbers of TM helices being predicted for the same protein.

With regard to secretion pathways other than Sec and Tat, the respective sorting signals are often poorly known, limited in numbers, or present only in a single mature protein. This hampers the development of dedicated prediction tools. Few notable exceptions are SecretomeP[60], SecretP[104] and NClassG+[105], of which only SecretomeP was still available at the time of writing of this thesis. The three tools are based on features extracted from sequences of non-classically secreted proteins, which often leads to biased or false results, as exemplified by the inclusion of Sec-secreted proteins in the outputs. Lastly, these tools are not able to determine which possible secretion pathway a protein of interest (POI) follows, but only indicate a higher probability to localise it in the extracellular milieu. Although this is a difficult task, different approaches may drastically improve the prediction of non-classically secreted proteins with precision. As demonstrated by the studies presented in **Chapter 2** of this thesis, starting from the actual knowledge of sorting signals rather than the available experimental data, proves to yield superior SCL predictions. In this regard, it is in fact possible to detect specific protein signatures and motifs, as deposited in online databases, with dedicated scanning tools, such as Interpro[94], Pfam[57], PROSITE[95] and TIGRfam[96]. Only Interpro IDs are listed here, but all of the respective databases in fact possess dedicated entries for the cleavage sites of fimbriae and pilins (IPR012902), T7SS-SPs (IPR010310, IPR041275, IPR000084, IPR006829, IPR000030), non-classical SPs (IPR005877, IPR022263, IPR023833), and a long list of other useful signatures. Despite being often very specific and restricted to a few proteins, or to small classes of proteins, these entries are very helpful, not only in the detection of non-classically secreted proteins, but also in identifying the protein secretion pathway followed by a particular protein.

Similarly, a large number of entries exists for the many known CW retention signals. This involves both the covalently and non-covalently CW-bound proteins, and the respective entries can be exploited to predict CW-associated proteins. Of note, a tool for CW anchor prediction called CW-PRED[106] has been developed, but it is restricted to

covalently attached proteins, limiting its applicability.

SCL prediction

In addition to single feature predictors, e.g. for SPs, TM helices or other sorting signals, there is a more comprehensive type of prediction, namely the prediction of the final subcellular compartment where a specific protein is localized. This SCL prediction is a key element within the functional annotation of a protein and it is of interest for multiple reasons, ranging from basic scientific knowledge to medical and industrial applications. For Gram-positive bacteria, there are four main localizations, i.e. the cytosol, plasma membrane, CW, and extracellular milieu. The aforementioned inner wall zone is a somewhat debated compartment that was, so far, never included in any SCL predictor. In this respect, it must be noted that, often, the assignment of a SCL to a protein can actually be a semantic issue[11]. Generally, to assign an SCL to a protein, three main approaches are available (extensively reviewed in [81,107–109]), which are based on the information exploited for the prediction: 1) the physico-chemical properties of the protein; 2) the presence of detectable sorting signals; and 3) the homology with known proteins and the subsequent transfer of an SCL designation. While the first approach is nowadays considered sub-optimal (at least if not combined with other approaches), the signal-based and the homology-based approaches are most extensively exploited. These two latter approaches have their respective advantages and disadvantages as discussed in detail in **Chapter 2** of this thesis. The signal-based approach has some particular benefits, the main one being that it can simulate or reproduce the same sorting signal-based mechanisms that would happen within a bacterial cell. On the other hand, this approach is limited by our current knowledge and understanding of sorting pathways and signals, as well as by the availability of tools for their detection. Such limitations should always be taken into account, though not to penalize the signal-based SCL prediction approach but, rather, to foster research in the direction of detecting and understanding all protein sorting pathways and the respective mechanisms. The knowledge thus generated can subsequently be applied to develop appropriate tools for improved SCL prediction.

Lastly, it has turned out convenient to combine different tools that are either able to detect the same type of signal, e.g. by combining SignalP, Phobius and PrediSi for the detection of SPs, or that will detect different types of signals, e.g. TatP and LipoP. With the first approach a compensation for each tool's possible downsides can be achieved, while the latter approach may lead to a reduction of false SCL predictions. Generally, it has been demonstrated that meta-approaches, based on the exploitation of multiple tools, even of multiple SCL predictors based on different approaches, leads to improved predictions, provided that the weights of individual predictions are assigned properly. This type of approach has been implemented in **Chapter 2** of this thesis, underscoring the superiority of meta-predictors for the genome-wide SCL of proteins.

SP efficiency prediction

One aspect of SP predictions that deserves particular attention, is the prediction of protein secretion efficiency as directed by a specific SP sequence, which was so far not possible. This problem is not merely of scientific interest, but it is also of high industrial relevance, because the costs of production of heterologous recombinant proteins are related to the efficiency of their secretion. In an industrial context, efficient secretion of a protein will be directly mirrored in the respective yields from the fermentation broth. For this reason, much effort has been attributed to understanding and removing various bottlenecks that reduce or hamper protein secretion at the industrial scale, and to understanding which SP sequence best drives the secretion of particular wild-type or recombinant POIs.

At present the most frequently used approach involves the screening of an SP library fused to the POI. This is known to be a very expensive and redundant approach and, to make matters worse, it is necessary to repeat this operation for each individual POI and each production host due to the lack of an adequate theoretical understanding of the underlying principles. While many of the relevant SP features that impact on protein secretion efficiency are known[68], there is no general model available to either predict or explain the resulting secretion level. This is even more so a challenge in the context of different POIs. Despite the fact that the SP has been known for 30 years, only the latest approaches that combine big datasets with ML are able to shine a little bit of light on the features that determine protein secretion efficiency[71], and to achieve predictability of the best possible SP-POI match[72]. Therefore, an approach combining high-throughput screening and ML was implemented in the studies described in **Chapter 5** of this thesis.

Scope of the thesis

In **Chapter 1** of this thesis, a brief introduction on the known bacterial protein secretion pathways and the respective mechanisms is presented, with a special focus on monoderm Gram-positive bacteria. Additionally, the scientific and biotechnological implications of protein secretion, and the relevant bioinformatics prediction tools for protein SCL are discussed.

Chapter 2 presents the GP^4 signal-based meta-predictor to determine protein SCLs in Gram-positive bacteria. Meta-predictors have long been known to outperform individual tools for SP or SCL predictions due to their ability to balance the strengths and weaknesses of their individual constituents. GP^4 represents the first implementation of such a signal-based approach for Firmicutes and Actinobacteria, and it is made available as a simple webserver. Difficulties in the construction of meta-predictors like GP^4 originate from the lack of stand-alone versions of the software components, and lack of standardized and easily programmatically parsable outputs. Additionally, the proposed GP^4 approach leaves

space for re-interpretation of the results in the light of an appropriate biological context, e.g. the particular species for which protein SCLs are predicted and its relationship to well-characterized model organisms, or the particular types of analysed proteins (i.e. wild-type or recombinant). A benchmark analysis proved GP[4] to be superior over other widely used SCL prediction tools, both in terms of accuracy of the prediction, as well as quality of the details provided on the sorting pathway(s) accessed by particular proteins.

Chapter 3 reviews the association between *Porphyromonas gingivalis*, a Gram-negative oral pathogen, and rheumatoid arthritis from a molecular perspective. Specifically, either secreted or outer-membrane-bound proteins of *P. gingivalis* seem to play a major role in the etiology of the disease and the generation of auto-antibodies. For this reason, an organism-specific signal-based SCL meta-predictor was developed, particularly taking into account the recently discovered type IX secretion system (T9SS), also called porin secretion system (PorSS), as well as possible non-canonical secretion pathways. By "mapping" all *P. gingivalis* proteins based on their SCL, a better understanding of known, novel, and putative virulence factors was achieved with implications for the identification of potentially druggable targets and inhibitors of virulence.

Chapter 4 of this thesis describes the comparison of community-associated (CA) and hospital-associated (HA) methicillin-resistant *S. aureus* (MRSA) aimed at identifying molecular traits that can be used to separate them. A comparative genomic and proteomic analysis was performed, showing that HA-MRSA and CA-MRSA have two different exproteome profiles. In this regard, a signal-based meta-prediction pipeline was developed to assign SCLs to all proteins detected in the exoproteome. This method was implemented in a way that only proteins with a SP would be classified as secreted, showing extremely high levels of non-classical secretion, either due to unknown, and thus unpredictable secretion mechanisms, or to other unspecific mechanisms. Intriguingly, proteins predicted to be non-classically secreted had a potentially relevant role in virulence and the epidemiological behaviour of the investigated MRSA isolates.

Chapter 5 presents a completely novel synthetic biology approach to understand and predict the secretion efficiency of a specific SP sequence. By combining high-throughput screening with a ML model and its interpretation, it was possible to perform the first round of a Design-Build-Test-Learn (DBTL) cycle aimed at completely elucidating the main SP characteristics. To this end, a library of approximately 12,000 SPs was screened for their efficiency in directing the secretion of an α-amylase by *B. subtilis*. The resulting data was used to train a Random Forest (RF) model, which in turn was interpreted with the game theory approach SHAP (SHapley Additive exPlanations). Subsequently, the model was used to modify SP sequences in order to obtain α-amylase secretion at a desired

level. Additionally, a library of pseudo-randomly designed SPs was *in silico* screened for high performing SPs. Out of the 21 tested pseudo-randomly designed SPs, 7 proved to be equal or superior to the wild-type. This study is of prime importance as it is now possible for the first time to generate an accurate predictive model of SP efficiency. In addition, the study provides important explanations on the general and specific SP characteristics that have a relevant impact on protein secretion efficiency.

The last experimental **Chapter 6** of this thesis illustrates an attempt to elucidate the possible role of Pro-peptides (Pros) in protein secretion, and their application potential. The results show that the contribution of the investigated Pros to protein secretion is, most likely, related to the characteristics of the specific sequences, rather than to their function as Pros. Additionally, by analyzing the growth media of protease-proficient and -deficient strains of *B. subtilis* by mass spectrometry, the proteases involved in the cleavage of specific Pros and the respective cleavage sites were investigated.

Finally, **Chapter 7** summarizes the overall findings outlined in this thesis, and places them within a broader context of the protein secretion field and its industrial applications. Possible improvements of the presented approaches are pointed out so that they may be adopted as standards for future fundamental investigations and the production of high-value proteins.

References

1. Verma, A., Rastogi, S., Agrahari, S. & Singh, A. Biotechnology in the realm of history. *J. Pharm. Bioallied Sci.* 3, 321 (2011).

2. Arranz-Otaegui, A., Gonzalez Carretero, L., Ramsey, M. N., Fuller, D. Q. & Richter, T. Archaeobotanical evidence reveals the origins of bread 14,400 years ago in northeastern Jordan. *Proc. Natl. Acad. Sci. U. S. A.* 115, 7925–7930 (2018).

3. Liu, L. *et al.* Fermented beverage and food storage in 13,000 y-old stone mortars at Raqefet Cave, Israel: Investigating Natufian ritual feasting. *J. Archaeol. Sci. Reports* 21, 783–793 (2018).

4. McGovern, P. *et al.* Early Neolithic wine of Georgia in the South Caucasus. *Proc. Natl. Acad. Sci.* 114, E10309–E10318 (2017).

5. McGovern, P. E. *et al.* Fermented beverages of pre- and proto-historic China. *Proc. Natl. Acad. Sci.* 101, 17593–17598 (2004).

6. Flamm, E. L. How FDA Approved Chymosin: A Case History. *Nat. Biotechnol.* 9, 349–351 (1991).

7. Arbige, M. V., Shetty, J. K. & Chotani, G. K. Industrial Enzymology: The Next Chapter. *Trends Biotechnol.* 37, 1355–1366 (2019).

8. Singh, R., Kumar, M., Mittal, A. & Mehta, P. K. Microbial enzymes: industrial progress in 21st century. *3 Biotech* 6, 174 (2016).

9. Ebert, M. C. & Pelletier, J. N. Computational tools for enzyme improvement: why everyone can – and should – use them. *Curr. Opin. Chem. Biol.* 37, 89–96 (2017).

10. Medema, M. H., van Raaphorst, R., Takano, E. & Breitling, R. Computational tools for the synthetic design of biochemical pathways. *Nat. Rev. Microbiol.* 10, 191–202 (2012).

11. Desvaux, M., Hébraud, M., Talon, R. & Henderson, I. R. Secretion and subcellular localizations of bacterial proteins: a semantic awareness issue. *Trends Microbiol.* 17, 139–145 (2009).

12. Coico, R. Gram staining. *Curr. Protoc. Microbiol.* Appendix 3, Appendix 3C (2005).

13. Megrian, D., Taib, N., Witwinowski, J., Beloin, C. & Gribaldo, S. One or two membranes? Diderm Firmicutes challenge the Gram-positive/Gram-negative divide. *Mol. Microbiol.* 113, 659–671 (2020).

14. Gibbons, N. E. & Murray, R. G. E. Proposals Concerning the Higher Taxa of Bacteria. *Int. J. Syst. Bacteriol.* 28, 1–6 (1978).

15. Tjalsma, H. *et al.* Proteomics of Protein Secretion by Bacillus subtilis : Separating the 'Secrets' of the Secretome. *Microbiol. Mol. Biol. Rev. 2* 68, 207–233 (2004).

16. Grassly, N. C. & Fraser, C. Mathematical models of infectious disease transmission. *Nat. Rev. Microbiol.* 6, 477–487 (2008).

17. Kuhn, A., Koch, H.-G. & Dalbey, R. E. Targeting and Insertion of Membrane Proteins. *EcoSal Plus* 7, (2017).

18. Kumazaki, K. *et al.* Structural basis of Sec-independent membrane protein insertion by

YidC. *Nature* 509, 516–520 (2014).
19. Harwood, C. R. & Cranenburgh, R. Bacillus protein secretion: an unfolding story. *Trends Microbiol.* 16, 73–9 (2008).
20. Saraogi, I. & Shan, S. Co-translational protein targeting to the bacterial membrane. *Biochim. Biophys. Acta - Mol. Cell Res.* 1843, 1433–1441 (2014).
21. Zhang, K., Su, L. & Wu, J. Recent Advances in Recombinant Protein Production by *Bacillus subtilis*. *Annu. Rev. Food Sci. Technol.* 11, 295–318 (2020).
22. Ling Lin Fu *et al.* Protein secretion pathways in Bacillus subtilis: Implication for optimization of heterologous protein secretion. *Biotechnol. Adv.* 25, 1–12 (2007).
23. Vörös, A. *et al.* SecDF as Part of the Sec-Translocase Facilitates Efficient Secretion of Bacillus cereus Toxins and Cell Wall-Associated Proteins. *PLoS One* 9, e103326 (2014).
24. Bolhuis, A. *et al.* SecDF of Bacillus subtilis , a Molecular Siamese Twin Required for the Efficient Secretion of Proteins. *J. Biol. Chem.* 273, 21217–21224 (1998).
25. Tjalsma, H., Bron, S. & van Dijl, J. M. Complementary impact of paralogous Oxa1-like proteins of Bacillus subtilis on post-translocational stages in protein secretion. *J. Biol. Chem.* 278, 15622–32 (2003).
26. Zweers, J. C. *et al.* Towards the development of Bacillus subtilis as a cell factory for membrane proteins and protein complexes. *Microb. Cell Fact.* 7, 10 (2008).
27. Dalbey, R. E., Kuhn, A., Zhu, L. & Kiefer, D. The membrane insertase YidC. *Biochim. Biophys. Acta - Mol. Cell Res.* 1843, 1489–1496 (2014).
28. Tjalsma, H., Bron, S. & Van Dijl, J. M. Complementary impact of paralogous Oxa1-like proteins of Bacillus subtilis on post-translocational stages in protein secretion. *J. Biol. Chem.* 278, 15622–15632 (2003).
29. He, H., Kuhn, A. & Dalbey, R. E. Tracking the Stepwise Movement of a Membrane-inserting Protein In Vivo. *J. Mol. Biol.* 432, 484–496 (2020).
30. Tjalsma, H., Bolhuis, A., Jongbloed, J. D., Bron, S. & van Dijl, J. M. Signal peptide-dependent protein transport in Bacillus subtilis: a genome-based survey of the secretome. *Microbiol. Mol. Biol. Rev.* 64, 515–47 (2000).
31. de Keyzer, J., van der Does, C. & Driessen, A. J. M. The bacterial translocase: a dynamic protein channel complex. *Cell. Mol. Life Sci.* 60, 2034–2052 (2003).
32. Collinson, I. The Dynamic ATP-Driven Mechanism of Bacterial Protein Translocation and the Critical Role of Phospholipids. *Front. Microbiol.* 10, 1217 (2019).
33. Traag, B. A., Pugliese, A., Setlow, B., Setlow, P. & Losick, R. A conserved ClpP-like protease involved in spore outgrowth in Bacillus subtilis. *Mol. Microbiol.* 90, 160–6 (2013).
34. Saito, A. *et al.* Post-liberation cleavage of signal peptides is catalyzed by the site-2 protease (S2P) in bacteria. *Proc. Natl. Acad. Sci.* 108, 13740–13745 (2011).
35. Henriques, G. *et al.* SppI Forms a Membrane Protein Complex with SppA and Inhibits Its Protease Activity in Bacillus subtilis. *mSphere* 5, e00724-20 (2020).
36. Feltcher, M. E. & Braunstein, M. Emerging themes in SecA2-mediated protein export. *Nat. Rev. Microbiol.* 10, 779–789 (2012).

37. Prabudiansyah, I. & Driessen, A. J. M. The canonical and accessory sec system of gram-positive bacteria. in *Current Topics in Microbiology and Immunology* vol. 404 45–67 (Springer Verlag, 2017).
38. Braunstein, M., Bensing, B. A. & Sullam, P. M. The Two Distinct Types of SecA2-Dependent Export Systems. in *Protein Secretion in Bacteria* 29–41 (ASM Press, 2019).
39. Bensing, B. A., Seepersaud, R., Yen, Y. T. & Sullam, P. M. Selective transport by SecA2: An expanding family of customized motor proteins. *Biochim. Biophys. Acta - Mol. Cell Res.* 1843, 1674–1686 (2014).
40. Tjalsma, H. & Van Dijl, J. M. Proteomics-based consensus prediction of protein retention in a bacterial membrane. *Proteomics* 5, 4472–4482 (2005).
41. Gardiner, J. H. *et al.* Lipoprotein N -Acylation in Staphylococcus aureus Is Catalyzed by a Two-Component Acyl Transferase System. *MBio* 11, 1–18 (2020).
42. Okuda, S. & Tokuda, H. Lipoprotein sorting in bacteria. *Annu. Rev. Microbiol.* 65, 239–259 (2011).
43. Hutchings, M. I., Palmer, T., Harrington, D. J. & Sutcliffe, I. C. Lipoprotein biogenesis in Gram-positive bacteria: knowing when to hold 'em, knowing when to fold 'em. *Trends Microbiol.* 17, 13–21 (2009).
44. Sibbald, M. J. J. B. *et al.* Mapping the Pathways to StaphylococcalPathogenesis by Comparative Secretomics. *Microbiol. Mol. Biol. Rev.* 70, 755–788 (2006).
45. Widdick, D. A. *et al.* The twin-arginine translocation pathway is a major route of protein export in Streptomyces coelicolor. *Proc. Natl. Acad. Sci.* 103, 17927–17932 (2006).
46. Goosens, V. J., Monteferrante, C. G. & Van Dijl, J. M. The Tat system of Gram-positive bacteria. *Biochim. Biophys. Acta - Mol. Cell Res.* 1843, 1698–1706 (2014).
47. Goosens, V. J. & van Dijl, J. M. Twin-Arginine Protein Translocation. in *Current Topics in Microbiology and Immunology* vol. 404 69–94 (Springer Verlag, 2016).
48. Frain, K. M., Robinson, C. & van Dijl, J. M. Transport of Folded Proteins by the Tat System. *Protein J.* 38, 377–388 (2019).
49. Bernal-Cabas, M. *et al.* Functional association of the stress-responsive LiaH protein and the minimal TatAyCy protein translocase in Bacillus subtilis. *Biochim. Biophys. Acta - Mol. Cell Res.* 1867, (2020).
50. Goosens, V. J., De-San-Eustaquio-Campillo, A., Carballido-López, R. & van Dijl, J. M. A Tat ménage à trois — The role of Bacillus subtilis TatAc in twin-arginine protein translocation. *Biochim. Biophys. Acta - Mol. Cell Res.* 1853, 2745–2753 (2015).
51. Ates, L. S., Houben, E. N. G. & Bitter, W. Type VII Secretion: A Highly Versatile Secretion System. *Microbiol. Spectr.* 4, 1–21 (2016).
52. Desvaux, M., Hébraud, M., Talon, R. & Henderson, I. R. Outer membrane translocation: numerical protein secretion nomenclature in question in mycobacteria. *Trends Microbiol.* 17, 338–40 (2009).
53. Unnikrishnan, M., Constantinidou, C., Palmer, T. & Pallen, M. J. The Enigmatic Esx

Proteins: Looking Beyond Mycobacteria. *Trends Microbiol.* 25, 192–204 (2017).
54. Sysoeva, T. A., Zepeda-Rivera, M. A., Huppert, L. A. & Burton, B. M. Dimer recognition and secretion by the ESX secretion system in Bacillus subtilis. *Proc. Natl. Acad. Sci. U. S. A.* 111, 7653–7658 (2014).
55. Pallen, M. J. The ESAT-6/WXG100 superfamily – and a new Gram-positive secretion system? *Trends Microbiol.* 10, 209–212 (2002).
56. Sutcliffe, I. C. New insights into the distribution of WXG100 protein secretion systems. *Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol.* 99, 127–131 (2011).
57. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* 47, D427–D432 (2019).
58. Poulsen, C., Panjikar, S., Holton, S. J., Wilmanns, M. & Song, Y. H. WXG100 protein superfamily consists of three subfamilies and exhibits an α-helical C-terminal conserved residue pattern. *PLoS One* 9, e89313 (2014).
59. Calvo, R. A. & Kearns, D. B. FlgM Is secreted by the flagellar export apparatus in Bacillus subtilis. *J. Bacteriol.* 197, 81–91 (2015).
60. Wang, G. *et al.* How are the Non-classically Secreted Bacterial Proteins Released into the Extracellular Milieu? *Curr. Microbiol.* 67, 688–695 (2013).
61. Kang, Q. & Zhang, D. Principle and potential applications of the non-classical protein secretory pathway in bacteria. *Appl. Microbiol. Biotechnol.* 104, 953–965 (2020).
62. Zhao, X. *et al.* Exoproteome Heterogeneity among Closely Related Staphylococcus aureus t437 Isolates and Possible Implications for Virulence. *J. Proteome Res.* 18, 2859–2874 (2019).
63. Ebner, P. & Götz, F. Bacterial Excretion of Cytoplasmic Proteins (ECP): Occurrence, Mechanism, and Function. *Trends Microbiol.* 27, 176–187 (2019).
64. Krishnappa, L. *et al.* Extracytoplasmic Proteases Determining the Cleavage and Release of Secreted Proteins, Lipoproteins, and Membrane Proteins in Bacillus subtilis. *J. Proteome Res.* 12, 4101–4110 (2013).
65. Desvaux, M. & Hébraud, M. Analysis of cell envelope proteins. in *Handbook of Listeria Monocytogenes* 359–393 (CRC Press, 2008).
66. Desvaux, M., Dumas, E., Chafsey, I. & HÃ©braud, M. Protein cell surface display in Gram-positive bacteria: from single protein to macromolecular protein structure. *FEMS Microbiol. Lett.* 256, 1–15 (2006).
67. Siegel, S. D., Reardon, M. E. & Ton-That, H. Anchoring of LPXTG-Like Proteins to the Gram-Positive Cell Wall Envelope. in *Current Topics in Microbiology and Immunology* vol. 404 159–175 (Springer Verlag, 2016).
68. Owji, H., Nezafat, N., Negahdaripour, M., Hajiebrahimi, A. & Ghasemi, Y. A comprehensive review of signal peptides: Structure, roles, and applications. *Eur. J. Cell Biol.* 97, 422–441 (2018).
69. Brockmeier, U. *et al.* Systematic Screening of All Signal Peptides from Bacillus subtilis:

A Powerful Strategy in Optimizing Heterologous Protein Secretion in Gram-positive Bacteria. *J. Mol. Biol.* 362, 393–402 (2006).

70. Degering, C. *et al.* Optimization of Protease Secretion in Bacillus subtilis and Bacillus licheniformis by Screening of Homologous and Heterologous Signal Peptides. *Appl. Environ. Microbiol.* 76, 6370–6376 (2010).

71. Peng, C. *et al.* Factors Influencing Recombinant Protein Secretion Efficiency in Gram-Positive Bacteria: Signal Peptide and Beyond. *Front. Bioeng. Biotechnol.* 7, 139 (2019).

72. Wu, Z. *et al.* Signal Peptides Generated by Attention-Based Neural Networks. *ACS Synth. Biol.* 9, 2154–2161 (2020).

73. Bogsch, E., Brink, S. & Robinson, C. Pathway specificity for a ΔpH-dependent precursor thylakoid lumen protein is governed by a 'Sec-avoidance' motif in the transfer peptide and a 'Sec-incompatible' mature protein. *EMBO J.* 16, 3851–3859 (1997).

74. Houben, E. N. G., Korotkov, K. V. & Bitter, W. Take five - Type VII secretion systems of Mycobacteria. *Biochim. Biophys. Acta - Mol. Cell Res.* 1843, 1707–1716 (2014).

75. Daleke, M. H. *et al.* General secretion signal for the mycobacterial type VII secretion pathway. *Proc. Natl. Acad. Sci. U. S. A.* 109, 11342–11347 (2012).

76. Shinde, U. & Inouye, M. Propeptide-mediated folding in subtilisin: The intramolecular chaperone concept. *Adv. Exp. Med. Biol.* 379, 147–154 (1996).

77. Demidyuk, I. V., Shubin, A. V., Gasanov, E. V. & Kostrov, S. V. Propeptides as modulators of functional activity of proteases. *Biomol. Concepts* 1, 305–322 (2010).

78. Kakeshita, H., Kageyama, Y., Ara, K., Ozaki, K. & Nakamura, K. Propeptide of Bacillus subtilis amylase enhances extracellular production of human interferon-α in Bacillus subtilis. *Appl. Microbiol. Biotechnol.* 89, 1509–17 (2011).

79. Kouwen, T. R. H. M. *et al.* Contributions of the Pre- And Pro-Regions of a Staphylococcus hyicus Lipase to Secretion of a Heterologous Protein by Bacillus subtilis. *Appl. Environ. Microbiol.* 76, 659–669 (2010).

80. Tian, P. & Bernstein, H. D. Identification of a post-targeting step required for efficient cotranslational translocation of proteins across the Escherichia coli inner membrane. *J. Biol. Chem.* 284, 11396–11404 (2009).

81. Nielsen, H., Tsirigos, K. D., Brunak, S. & von Heijne, G. A Brief History of Protein Sorting Prediction. *Protein J.* 38, 200–216 (2019).

82. McGeoch, D. J. On the predictive recognition of signal peptide sequences. *Virus Res.* 3, 271–286 (1985).

83. Almagro Armenteros, J. J. *et al.* SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* 37, 420–423 (2019).

84. Savojardo, C., Martelli, P. L., Fariselli, P. & Casadio, R. DeepSig: Deep learning improves signal peptide detection in proteins. *Bioinformatics* 34, 1690–1696 (2018).

85. Hiller, K., Grote, A., Scheer, M., Münch, R. & Jahn, D. PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Res.* 32, W375-9 (2004).

86. Käll, L., Krogh, A. & Sonnhammer, E. L. L. A Combined Transmembrane Topology and Signal Peptide Prediction Method. *J. Mol. Biol.* 338, 1027–1036 (2004).
87. Fariselli, P., Finocchiaro, G. & Casadio, R. SPEPlip: the detection of signal peptide and lipoprotein cleavage sites. *Bioinformatics* 19, 2498–2499 (2003).
88. Bendtsen, J. D., Nielsen, H., Widdick, D., Palmer, T. & Brunak, S. Prediction of twin-arginine signal peptides. *BMC Bioinformatics* 6, 167 (2005).
89. Rose, R. W., Brüser, T., Kissinger, J. C. & Pohlschröder, M. Adaptation of protein secretion to extremely high-salt conditions by extensive use of the twin-arginine translocation pathway. *Mol. Microbiol.* 45, 943–950 (2002).
90. Dilks, K., Rose, R. W., Hartmann, E. & Pohlschröder, M. Prokaryotic utilization of the twin-arginine translocation pathway: A genomic survey. *J. Bacteriol.* 185, 1478–1483 (2003).
91. Bagos, P. G., Nikolaou, E. P., Liakopoulos, T. D. & Tsirigos, K. D. Combined prediction of Tat and Sec signal peptides with hidden Markov models. *Bioinformatics* 26, 2811–2817 (2010).
92. Juncker, A. S. *et al.* Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci.* 12, 1652–1662 (2003).
93. Bagos, P. G., Tsirigos, K. D., Liakopoulos, T. D. & Hamodrakas, S. J. Prediction of lipoprotein signal peptides in Gram-positive bacteria with a Hidden Markov Model. *J. Proteome Res.* 7, 5082–5093 (2008).
94. Mitchell, A. L. *et al.* InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* 47, D351–D360 (2019).
95. Sigrist, C. J. A. *et al.* New and continuing developments at PROSITE. *Nucleic Acids Res.* 41, D344-7 (2013).
96. Haft, D. H. *et al.* TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res.* 29, 41–3 (2001).
97. Zuber, B. B. *et al.* Granular Layer in the Periplasmic Space of Gram-Positive Bacteria and Fine Structures of Enterococcus gallinarum and Streptococcus gordonii Septa Revealed by Cryo-Electron Microscopy of Vitreous Sections. *J. Bacteriol.* 188, 6652–6660 (2006).
98. Krogh, a, Larsson, B., von Heijne, G. & Sonnhammer, E. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580 (2001).
99. Tusnády, G. E. & Simon, I. The HMMTOP transmembrane topology prediction server. *Bioinformatics* 17, 849–850 (2001).
100. Jones, D. T. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* 23, 538–544 (2007).
101. Viklund, H. & Elofsson, A. OCTOPUS: Improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics* 24, 1662–1668 (2008).

102. Viklund, H., Bernsel, A., Skwark, M. & Elofsson, A. SPOCTOPUS: A combined predictor of signal peptides and membrane protein topology. *Bioinformatics* 24, 2928–2929 (2008).
103. Käll, L., Krogh, A. & Sonnhammer, E. L. L. An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics* 21 Suppl 1, i251-7 (2005).
104. Yu, L. *et al.* SecretP: Identifying bacterial secreted proteins by fusing new features into Chou's pseudo-amino acid composition. *J. Theor. Biol.* 267, 1–6 (2010).
105. Restrepo-Montoya, D., Pino, C., Nino, L. F., Patarroyo, M. E. & Patarroyo, M. A. NClassG+: A classifier for non-classically secreted Gram-positive bacterial proteins. *BMC Bioinformatics* 12, 21 (2011).
106. Litou, Z. I., Bagos, P. G., Tsirigos, K. D., Liakopoulos, T. D. & Hamodrakas, S. J. Prediction of cell wall sorting signals in gram-positive bacteria with a hidden markov model: Application to complete genomes. *J. Bioinform. Comput. Biol.* 6, 387–401 (2008).
107. Nielsen, H. Protein Sorting Prediction. in *Methods in Molecular Biology* vol. 1615 23–57 (Humana Press Inc., 2017).
108. Nielsen, H. Predicting Subcellular Localization of Proteins by Bioinformatic Algorithms. in *Current Topics in Microbiology and Immunology* vol. 404 129–158 (Springer Verlag, 2015).
109. Wan, S. & Mak, M.-W. *Machine Learning for Protein Subcellular Localization Prediction*. (DE GRUYTER, 2015).